# Abstract

Our intention is to find similarity among the time series expressions of the genes in microarray experiments. It is hypothesized that at a given time point the concentration of one gene's mRNA is directly affected by the concentration of other gene's mRNA, and may have biological significance. We define dissimilarity between two time-series data set as the variance of Euclidean distances of each time points. The large numbers of gene expressions make the calculation of variance of distance in each point computationally expensive and therefore computationally challenging in terms of execution time. For this reason we use autoregressive model which estimates nineteen points gene expression to a three point vector. It allows us to find variance of difference between two data sets without point-to-point matching. Previous analysis from the microarray experiments data found that 62 genes are regulated following EGF (Epidermal Growth Factor) and HRG (Heregulin) treatment of the MCF-7 breast cancer cells. We have chosen these suspected cancer-related genes as our reference and investigated which additional set of genes has similar time point expression profiles. Keeping variance of difference as a measure of distance, we have used several methods for clustering the gene expression data, such as our own maximum clique finding heuristics and hierarchical clustering. The results obtained were validated through a text mining study. New predictions from our study could be a basis for further investigations in the genesis of breast cancer. Overall in 84 new genes are found in which 57 genes are related to cancer among them 35 genes are associated with breast cancer.